WHAT IS CLAIMED IS

1. A method for reducing magnitudes of output traffic bursts in a streaming media cache comprises:

receiving a request from a first client system for a stream of media data, the stream of media data including a first streaming media data packet and a second streaming media data packet;

receiving a request from a second client system for the stream of media data;

receiving the first streaming media data packet from an upstream server, the first streaming media data packet including a delivery time;

determining a first modified delivery time for the first streaming media data packet;

determining a second modified delivery time for the first streaming media data packet, the first modified delivery time different from the second modified delivery time;

modifying the first streaming media data packet with the first modified delivery time to form a first modified first streaming media data packet;

modifying the first streaming media data packet with the second modified delivery time to form a second modified first streaming media data packet;

outputting the first modified first streaming media data packet to the first client system at the first modified delivery time; and

outputting the second modified first streaming media data packet to the second client system at the second modified delivery time.

2. The method of claim 1 wherein determining a first modified delivery time comprises adding a first delay value to the delivery time.

3. The method of claim 2 wherein the first delay value is selected from the range: 0 to approximately 500 milliseconds.

4. The method of claim 3 wherein the first delay value is pseudo-randomly selected from the range.

5. The method of claim 1 further comprising storing a payload portion of the first streaming media in a storage within the streaming media cache.

6. The method of claim 2 wherein the second streaming media data packet includes a delivery time, the method further comprising:

determining a first modified delivery time for the second streaming media data packet;

determining a second modified delivery time for the second streaming media data packet, the first modified delivery time different from the second modified delivery time;

modifying the second streaming media data packet with the first modified delivery time to form a first modified second streaming media data packet;

modifying the second streaming media data packet with the second modified delivery time to form a second modified second streaming media data packet;

outputting the first modified second streaming media data packet to the first client system at the first modified delivery time; and

outputting the second modified second streaming media data packet to the second client system at the second modified delivery time.

7. The method of claim 6 wherein determining the first modified delivery time for the second streaming media data packet comprises adding the first delay value to the delivery time of the second streaming media data packet.

8. The method of claim 1 further comprising:

receiving a data file from the upstream server, the data file including a payload portion of the first streaming media data packet and a payload portion of the second streaming media data packet; and

storing the data file in a storage within the streaming media cache.

9. A computer system for providing streaming media data to client systems with reduced magnitude traffic bursts comprises:

a first thread configured to receive a request from a first client system and a second client system for a stream of data packets, wherein the stream includes a first data packet and a second data packet, the first thread also configured to specify a first client delay and a second client delay;

a second thread coupled to the first thread, the second thread configured to receive the first data packet and the second data packet from an upstream server, wherein the first data packet specifies a first delivery time and the second data packet specifies a second

delivery time, the second thread also configured to form a first delayed first data packet in response to the first client delay and to form a second delayed first data packet in response to the second client delay, wherein the first delayed first data packet specifies a first delayed delivery time and the second delayed first data packet specifies a second delayed delivery time;

a third thread configured to receive the first delayed first data packet and to provide the first delayed first data packet to the first client system in response to the first delayed delivery time; and

a fourth thread configured to receive the second delayed first data packet and to provide the second delayed first data packet to the second client system in response to the second delayed delivery time.

10. The computer system of claim 9 wherein the second thread is configured to form the first delayed first data packet in response to the first client delay by adding the first client delay to the first delivery time.

11. The computer system of claim 10 wherein the first client delay is selected from the range: 0 to approximately 500 milliseconds.

12. The computer system of claim 9 further comprising a thread configured to store payload portions of the first data packet and payload portions of the second data packet in a memory

13. The computer system of claim 9 further

wherein the second thread is also configured to form a first delayed second data packet in response to the first client delay and to form a second delayed second data packet in response to the second client delay, wherein the first delayed second data packet specifies a first delayed delivery time and the second delayed second data packet specifies a second delayed delivery time;

wherein the third thread is also configured to receive the first delayed second data packet and to provide the first delayed second data packet to the first client system in response to the first delayed delivery time specified therein; and

wherein the fourth thread is also configured to receive the second delayed second data packet and to provide the second delayed second data packet to the second client system in response to the second delayed delivery time specified therein.

14. The computer system of claim 13 wherein the second thread is also configured to form the first delayed second data packet in response to the first client delay by adding the second client delay to the second delivery time.

15. The computer system of claim 14 wherein the second client delay is pseudo-randomly selected.

16. A method for reducing peak output traffic bursts in a computer system where a first packet of data is scheduled to be delivered to more than one downstream client system substantially simultaneously comprises:

delaying a packet delivery time for the first packet of data to be delivered to a first downstream client system; and

delaying the packet delivery time for the first packet of data to be delivered to a second downstream client system;

wherein the first packet of data is to be delivered to the first downstream client system at a time different than when the first packet of data is to be delivered to the second downstream client system.

17. The method of claim of claim 16 wherein the first packet of data is framed.

18. The method of claim 16 wherein the first packet of data comprises streaming media data.

19. The method of claim 16 wherein delaying the packet delivery time for the first packet of data to be delivered to the first downstream client system comprises delaying the first packet of data by a delay factor selected from 0-500 milliseconds.

20. The method of claim 19 further comprising delaying a packet delivery time for a second packet of data to be delivered to the first downstream client system by the delay factor.